L'Intelligenza Artificiale rappresenta le nostre emozioni attraverso il linguaggio

Uno studio della Fondazione per la Sostenibilità Digitale pubblicato dalla prestigiosa rivista scientifica JMIR spiega perché l'IA sembra capace di comprenderci ed evidenzia possibilità e limiti del suo impiego nella terapia psichiatrica

Scoperta la riflessività algoritmica: la forma in cui le persone rappresentano i loro stati mentali è integrata nella struttura della lingua, che la macchina riproduce perfettamente

La ricerca nasce nell'ambito dell'iniziativa Digital4Aid e del progetto mySMART Diary

Roma, 5 novembre 2026

La Fondazione per la Sostenibilità Digitale ha pubblicato su JMIR AI – la più prestigiosa rivista scientifica internazionale *peer-reviewed* dedicata alla ricerca sull'AI per la sanità digitale – uno studio che spiega **il meccanismo per cui l'intelligenza artificiale generativa ci appare in grado di comprendere le emozioni** e gli stati mentali delle persone.

Questo meccanismo si chiama "**riflessività algoritmica**" ed è la chiave per capire come usare l'IA in contesti terapeutici assistiti – sfruttandola come laboratorio di ricerca, palestra formativa e supporto all'analisi – senza incorrere nel rischio della "pseudo empatia", ovvero della convinzione di essere capiti da una macchina che in realtà non ha questa capacità, o in utilizzi impropri.

Un percorso alternativo per un esperimento mai tentato prima

Invece di domandarsi "se le macchine hanno una mente" (la domanda che ha dominato il dibattito negli ultimi anni) questo studio ha deciso di seguire una via alternativa: misurare **quanto dei meccanismi della nostra mente è presente nel funzionamento dei grandi modelli linguistici**, che non hanno per definizione una psiche - alcuna esperienza, alcuna emozione, alcuna intenzione.

In specifico, lo studio "*Algorithmic Reflectivity in Large Language Models: a Methodological Study of Linguistic Structure in Mentalization*" si concentra sulla **mentalizzazione**, ovvero la capacità delle persone di interpretare i comportamenti propri e altrui in termini di stati mentali (intenzioni, credenze, emozioni…) dotati di senso.

La mentalizzazione è una delle competenze fondative della mente umana, che la psicologia clinica ha sempre considerato radicata interamente nell' esperienza affettiva e relazionale: in questa visione, il linguaggio con cui si producono le interpretazioni è semplicemente un mezzo per osservare un processo psichico.

Attraverso l'analisi da parte di esperti, condotta su 50 dialoghi reali avvenuti tra una persona e un modello linguistico configurato come "intervistatore clinico", lo studio ha dimostrato, invece, che **il linguaggio** – ovvero ciò che le macchine sono in grado di riprodurre perfettamente, sotto forma di regolarità statistica – **contiene anche una sorta di "grammatica del pensiero" riflessivo**.

La lingua “contiene” la struttura e la forma del processo di mentalizzazione, che possono essere riprodotte da chiunque la padroneggi - uomo o macchina: è per questo che quando dialoghiamo con un’AI vediamo che essa sembra comprenderci. Ma questo è anche un meccanismo che ha una grande importanza per la capacità di utilizzare l’AI come supporto nella terapia clinica.

“*Il valore di questo studio è molteplice. Con il filtro della riflessività algoritmica, ciò che nella mente umana resta implicito diventa, per la prima volta, osservabile e misurabile in isolamento, tramite il linguaggio*” spiega **Stefano Epifani, Presidente della Fondazione per la Sostenibilità Digitale** “*Abbiamo trovato nei modelli linguistici la struttura della riflessione, elaborata in un modo convincente persino per gli specialisti: è un risultato che apre possibilità enormi per la ricerca e la clinica, ma dice anche quanto sia urgente governare questi strumenti, in un’ottica di sostenibilità digitale*” .

La grammatica del pensiero riflessivo

Lo studio è stato strutturato secondo un preciso dispositivo. Sono stati presi in considerazione cinquanta dialoghi in italiano svolti tra partecipanti umani e un grande modello linguistico, configurato come “intervistatore clinico”.

Al termine di ogni dialogo è stato chiesto al modello linguistico di generare un profilo di mentalizzazione strutturato secondo le indicazioni proprie dell’approccio clinico *Mentalization-Based Treatment* (MBT), uno dei più consolidati a livello internazionale, che si basa appunto sull’interpretazione degli stati mentali della persona.

I profili sono stati valutati (in cieco, quindi all’oscuro dell’origine dei testi) da cinque psichiatri certificati MBT con oltre dieci anni di esperienza clinica. Per evitare che il giudizio sulla validità dei profili si basasse solo sulla fluidità del testo, è stato introdotto un controllo negativo, condotto abbinando deliberatamente alcuni dialoghi a profili di mentalizzazione sbagliati, immediatamente riconosciuti e penalizzati dai valutatori.

Il verdetto degli specialisti clinici sui dialoghi analizzati è stato netto: i **profili generati dalla macchina hanno mostrato una struttura riflessiva coerente e clinicamente interpretabile, con un forte accordo tra i valutatori**. Il modello linguistico si è dimostrato particolarmente efficace proprio nelle operazioni che costituiscono l’ossatura di un discorso riflessivo: rendere espliciti i nessi tra emozione, cognizione e comportamento; distinguere la prospettiva di sé da quella dell’altro; sintetizzare in una interpretazione organizzata un materiale emotivo eterogeneo.

Prospettive: l'IA nei contesti terapeutici assistiti

Se la forma della mentalizzazione “abita” nel linguaggio, il dialogo con il modello linguistico può entrare nei contesti della psicoterapia e della psicoanalisi per assisterla.

L’IA può essere usata come “**laboratorio di ricerca**”, perché il funzionamento riflessivo, finora osservabile solo immerso nell'esperienza soggettiva, diventa isolabile, misurabile, confrontabile; può servire da **palestra formativa** per i clinici, usando il modello linguistico per generare casi, varianti narrative e scenari controllati su cui allenare il giudizio clinico, anche per le condizioni più difficili; è un **supporto all’analisi**, capace di rendere visibile la struttura del discorso riflessivo nei testi dei pazienti, liberando tempo e risorse cognitive del terapeuta.

I rischi: quando la forma della comprensione basta a convincere

La stessa scoperta che apre queste prospettive ne definisce anche i rischi. Il primo è la **pseudo empatia**: un linguaggio che ha la forma perfetta della comprensione produce la sensazione di essere capiti, senza che nessuno stia davvero capendo. La fiducia che rende efficace una terapia, ricorda lo studio, non è una proprietà del sistema ma un fenomeno relazionale, fondato sul corpo, sulla contingenza affettiva, sulla capacità di tollerare l'incertezza.

Il secondo è la **profilazione psicologica**: una macchina che produce interpretazioni convincenti per gli specialisti è una macchina potenzialmente capace di interpretarci tutti. Lo studio mostra che perfino il giudizio clinico esperto risponde alla forma linguistica della riflessione, indipendentemente dal fatto che dietro ci sia un'esperienza reale.

Il terzo è la **compiacenza algoritmica**: la tendenza documentata dei modelli ad assecondare l'interlocutore può amplificare convinzioni rigide e narrazioni disfunzionali proprio nelle persone più vulnerabili.

Da qui la conclusione operativa: l'IA va intesa come **strumento dentro la relazione di cura, sotto supervisione professionale e mai come suo sostituto**.

Dalla ricerca alla cura: le radici in Digital4Aid e mySMART Diary

Lo studio pubblicato su JIMR AI è nato nell'ambito di Digital4Aid, l'iniziativa con cui la Fondazione per la Sostenibilità Digitale e la sua community di partner identificano, progettano e donano alla comunità progetti digitali al servizio degli obiettivi dell'Agenda 2030 — dall'inclusione alla tutela della salute delle categorie più fragili.

In particolare, la ricerca ha coinvolto anche l'ecosistema scientifico alla base di *mySMART Diary*, primo progetto frutto dell'iniziativa e prima web app a livello internazionale pensata per supportare i terapeuti nel trattamento dei disturbi del comportamento alimentare.

Il direttore scientifico dell'app, lo psichiatra e psicoterapeuta Giuliano Castigliego, è co-autore della ricerca, e i Servizi Psichiatrici del Canton Grigioni — sede di sperimentazione dell'app — sono rappresentati nel gruppo di studio. Un percorso coerente: prima costruire strumenti digitali per la cura, poi sottoporre a verifica scientifica rigorosa che cosa l'IA può e non può fare quando entra nei contesti clinici.

«Come clinico, trovo affascinante che una macchina possa riprodurre con tanta precisione alcune forme linguistiche della mentalizzazione" commenta **Giuliano Castigliego**. *"Ma proprio questo risultato ci aiuta anche a vedere meglio ciò che la macchina non possiede: l'esperienza dell'incontro con l'altro. In psicoterapia comprendere non significa soltanto formulare correttamente uno stato mentale; significa costruire insieme, nell'incertezza della relazione, un significato che prima non c'era. L'IA può aiutarci in questo processo, ma non può sostituirlo.»*

Lo studio "*Algorithmic Reflectivity in Large Language Models: a Methodological Study of Linguistic Structure in Mentalization*" è firmato da Stefano Epifani (Fondazione per la Sostenibilità Digitale, Università di Pavia), Giuliano Castigliego (Fondazione per la Sostenibilità Digitale, Accademia Psicoanalitica della Svizzera Italiana), Giuliano Razzicchia (Digital Transformation Institute), Elisabeth Seiwald-Sonderegger e Laura Kecskemeti (Servizi Psichiatrici del Canton Grigioni). Lo studio è disponibile su JMIR-AI.

Per ulteriori informazioni o approfondimenti, visitare il sito: www.sostenibilitadigitale.it

L'elenco dei Partner e delle Università che attualmente fanno parte della Fondazione può essere consultato al seguente link

Informazioni su Fondazione per la Sostenibilità Digitale

La Fondazione per la Sostenibilità Digitale è la prima Fondazione di Ricerca in Italia che analizza le correlazioni tra trasformazione digitale e sostenibilità con l'obiettivo di supportare istituzioni e imprese nella costruzione di un futuro migliore. La sua mission è quella di studiare le dinamiche indotte dalla trasformazione digitale, con particolare riferimento agli impatti sulla sostenibilità ambientale, culturale, sociale ed economica. In quest'ottica la Fondazione sviluppa attività di ricerca, fornisce letture ed interpretazioni della trasformazione digitale, offre

indicazioni operative per gli attori coinvolti, intercetta le tendenze del cambiamento e ne analizza gli impatti rispetto allo sviluppo sostenibile.
La Fondazione agisce attraverso una struttura costituita da esperti indipendenti, istituzioni, imprese e università. La Fondazione si ispira e si riconosce nel Manifesto per la Sostenibilità Digitale, una visione sul ruolo delle tecnologie per un futuro sostenibile, che definisce i princìpi sulla base dei quali propone di orientare lo sviluppo tecnologico per contribuire al "soddisfacimento dei bisogni della generazione presente senza compromettere la possibilità delle generazioni future di realizzare i propri".
Ai soci e partner della Fondazione si affianca la Rete delle Università che costituisce il sistema di competenze al quale fa riferimento la Fondazione per lo sviluppo dei suoi progetti e che rappresenta un esempio virtuoso di collaborazione tra istituzioni ed aziende nello sviluppo di progetti e di attività dedicati alla sostenibilità digitale. Tra le Università che fanno parte della Rete, l'Università Sapienza di Roma, l'Università di Pavia, l'Università degli Studi di Cagliari, l'Università degli Studi di Palermo, l'Università degli Studi di Firenze, l'Università degli Studi di Trieste, l'Università di Perugia, L'Università per Stranieri di Perugia, l'Università di Siena, l'Università degli Studi di Urbino Carlo Bo, l'Università degli Studi di Torino, l'Istituto di Studi Politici "S. Pio V", l'Università degli Studi di Sassari, l'Università della Tuscia, l'Università del Salento, l'Università di Macerata, l'Università di Trento, l'Università degli Studi di Ferrara, l'Università della Calabria.

Contatti per la stampa

Fondazione per la Sostenibilità Digitale

Vilma Bosticco
(vilma.bosticco@sostenibilitadigitale.it)
Federico Pesce
(federico.pesce@sostenibilitadigitale.it)

Prima Pagina Comunicazione

Caterina Ferrara (caterina@primapagina.it)